Noah Bhuta, Stella El-Fishawy, Alena Zeng
DATA 11900

Objective:

For this project, We are interested in predicting the type of crime committed based on specific geographical areas in Chicago and economic factors related to the location of the crime. We will then explore the relationships between the predictors and type of crime committed to deduce the root cause of why specific crimes might be committed given certain economic conditions. Furthermore, we are interested in exploring the most common crime, theft, more in depth. We will predict theft rate based on seven unique demographic variables and determine what demographic variables are more predictive of theft.

The Dataset:

For this project we used two datasets both from the Chicago City Data Portal. The first dataset reflects reported incidents of crime (with the exception of murders) that occurred in Chicago in 2015. The data is collected from the Chicago Police Department's Citizen Law Enforcement Analysis and Reporting (CLEAR). This dataset consists of 265,000 unique occurrences of crimes and 22 descriptors of the crime committed, including the type of the crime, a description of the location where it occurred, whether it resulted in arrest, the date and time it took place, and its geographical location [1]. The second dataset contains six socioeconomic indicators of poverty as well as a "hardship index" for the City of Chicago, for the years 2008 to 2012. This data is extracted from the U.S. Census Bureau. The dataset contains data for the 77 community areas in Chicago [2]. The third dataset contains information about the total population in each community area in 2015, extracted from the U.S. Census Bureau and American Community Survey, and compiled by the Heartland Alliance [3].

Literature Review:

Researchers across a wide variety of disciplines have used publicly available datasets on violent and property crimes in Chicago and other cities to understand the frequency and characteristics of crime and to predict the occurrence of future crime. For instance, through analyzing gun violence in Chicago, sociologists found that social network science, which describes the connections between people, organizations, and places, can help predict the probability of gunshot victimization, and better guide the deployment of intervention resources to potential locations of elevated violent activity [4]. In terms of forecasting, many models use a seismic approach to visualize and predict crime hotspots geographically on an area or grid based level either arbitrarily or by police beats. However, recent research using data on property crime in the UK, which uses kernel density estimation to calculate the risk level at a given location based on risk contributions from nearby preceding crimes, has shown that network based models on the street level can better take into account the social environment of the city and its networks for transportation and communication. These factors can influence the likelihood and location of crime, so taking them into account increases predictive accuracy [5].

In 2022, data and social scientists at UChicago developed a stochastic inference algorithm that forecasts future crime with 90% accuracy for crimes predicted per week within ~1,000 feet. Their model divides the city into spatial tiles rather than existing community or political boundaries to avoid bias, and it looks for patterns in the time and spatial coordinates of past occurrences before constructing a communicating network of local estimators to predict future infractions. Their predictions suggested that increased crime is biased by neighborhood socioeconomic status, draining policy resources from socioeconomically disadvantaged areas, as demonstrated in eight major US cities [6]. At the same time, their work has raised questions about the ethics of predictive crime models. These kinds of tools shouldn't be used to direct law enforcement, but rather serve as a model of urban environments that can show the impact of different variables to better guide urban policies as a whole.

EDA/Characteristics of the Sample:

In 2015, theft was the most frequently committed crime in Chicago with 57,350 thefts occurring throughout the year, followed closely by the 48,923 batteries committed as seen in *Figure 1.* The higher occurrence of specific types of crimes is something we will have to account for as we move on to machine learning.

Noah Bhuta, Stella El-Fishawy, Alena Zeng
DATA 11900

Figure 1

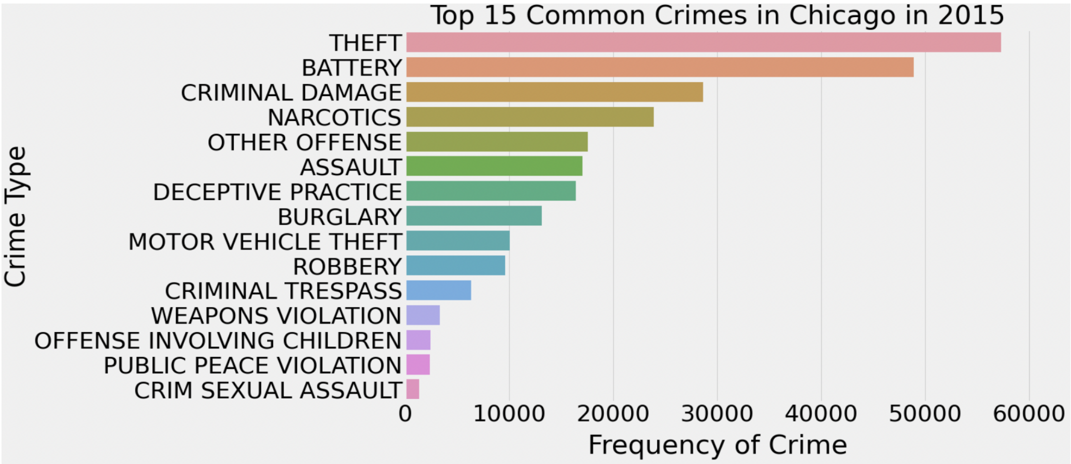
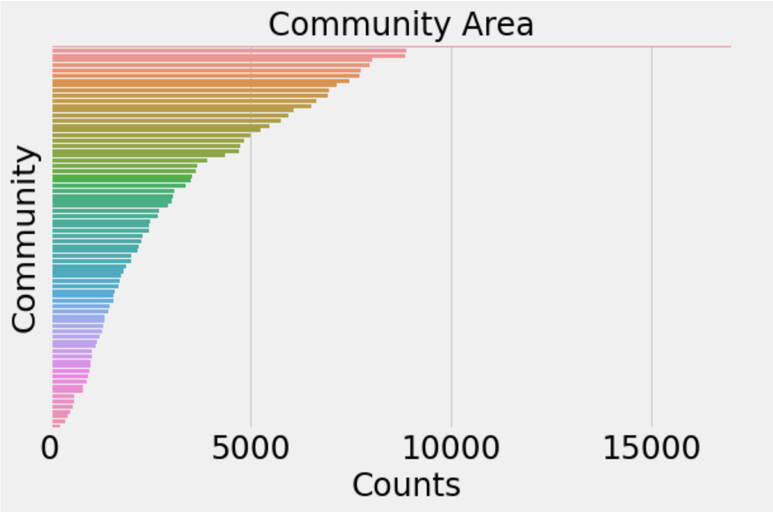


Figure 2 aggregates the amount of crime by community areas in Chicago. Chicago is divided into 77 community areas for planning and statistical purposes, and certain areas, such as Austin, Near North Side, and South Shore experience greater amounts of crime than others. * For aesthetic appeal, we removed the y labels to focus more on the overall distribution of total crimes per different area

Figure 2



Data Cleaning and Preparation:

Step 1. Merged our two datasets based on the community area column that both datasets had in common.

Step 2. Converted the date column to datetime format to enable time based analysis.

Step 3. Removed all rows that contained NaN values. There were not many in the dataset, therefore it was not a significant loss in data.

Step 4. Converted the categorical feature columns to new binary columns for each unique category in the original categorical column. Each row in the new columns will have a value of 1 if the category is present in the original column for that row, and 0 otherwise.

Step 5. Narrowed down the variety of crimes from our output variable from thirty-three to the nine most common ones and excluded the category “Other Offense”.

Step 6. Choose relevant features and engineered features for our model and split the data into input variables and an output variable. *More info on feature selection below

Step 7. Split the input and output data into training and test data using 80/20 split.

Step 8. Standardized the training and testing features.

Feature Selection:

When choosing features to predict the type of crime committed, it's important to consider variables that have been found to be relevant to criminal behavior in previous studies. In this case, we selected several demographic and socioeconomic variables that have been associated with crime rates, as well as one geographic variable. However, it's important to note that the relevance of these features may vary depending on the specific type of crime being committed. Some types of crimes may be more closely associated with certain demographic or socioeconomic factors than others. Below, we will examine how each of the features we have chosen may be relevant to predicting the type of crime committed:

1. PERCENT HOUSEHOLDS BELOW POVERTY: Poverty is a significant risk factor for criminal activity. Certain types of crimes, such as theft and robbery, may be more likely to occur in areas with high levels of poverty.

2. PERCENT AGED 16+ UNEMPLOYED: Unemployment may also be a significant risk factor for certain types of crime. For example, individuals who are unemployed may be more likely to engage in theft, burglary, or other property crimes.

3. PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA: A lack of education is also a risk factor for criminal activity. Individuals with lower levels of education may have fewer job opportunities and may be more likely to engage in criminal behavior to make ends meet.

4. PERCENT OF HOUSING CROWDED: Overcrowding in housing may be a risk factor for various types of crime, including burglary and theft.

5. PERCENT AGED UNDER 18 OR OVER 64: Certain types of crime, such as assault or battery, may be more likely to occur in areas with high percentages of young people. Conversely, crimes such as fraud or financial scams may be more likely to occur in areas with high percentages of older adults.

6. PER CAPITA INCOME: As previously noted, income is a significant risk factor for criminal activity. Certain types of crimes, such as theft or fraud, may be more likely to occur in areas with low per capita income.

7. HARDSHIP INDEX: The hardship index is a composite measure that takes into account several socioeconomic factors. Areas with high levels of hardship may be at higher risk for a variety of crimes.

8. DOMESTIC: Knowing whether the crime was Domestic or not is important because domestic related crimes can be associated with specific crimes. For example, people are less likely to steal from their family.

9. ARREST: Knowing whether the criminal was arrested or not can be an important factor in determining the severity of the crime. For more serious crimes, an individual has a higher chance of being arrested.

10. COMMUNITY AREA: Finally, including community area as a feature may be important for predicting the type of crime committed. Different types of crimes may be more likely to occur in different areas of the city, and accounting for these differences may improve the accuracy of the model.

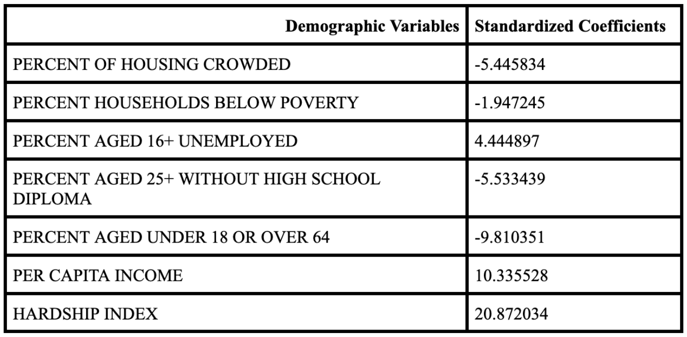
In summary, the features we chose are all relevant to predicting the type of crime committed. However, it's important to keep in mind that the relevance of these features may vary depending on the specific type of crime being committed. By understanding the relationships between these features and different types of crime, we can better develop strategies to prevent and address criminal activity in different communities.

Theft Prediction With Multiple Linear Regression:

To better understand how different economic factors affect the likelihood of crime, we conducted a multiple linear regression analysis to predict the number of thefts that occur in different Chicago community areas.

We first filtered the crime dataset for thefts, then aggregated the number of thefts by community area. Using the third dataset on community area populations in 2015, we standardized these counts as the number of thefts per 1000 people. Then, we merged this dataframe with the second dataset on socioeconomic factors (Features 1–7), standardized the columns, and used an 80/20 split to divide the community areas into training and testing data. After running the multiple linear regression using each socioeconomic variable, with a mean squared error of 116.27 and mean absolute error of 7.721, we found that hardship index is the strongest predictor for theft, followed by per capita income, shown below in *Figure* 3. The model equation obtained is $\textit{theft rate} = 21.139 - 5.446 * \textit{percent of housing crowded} - 1.947 * \textit{percent households below poverty} + 4.445 * \textit{percent aged 16+ unemployed} - 5.533 * \textit{percent aged 25+ without HS diploma} - 9.810 * \textit{percent aged under 18 or over 64} + 10.336 * \textit{per capita income} + 20.872 * \textit{hardship index}$.

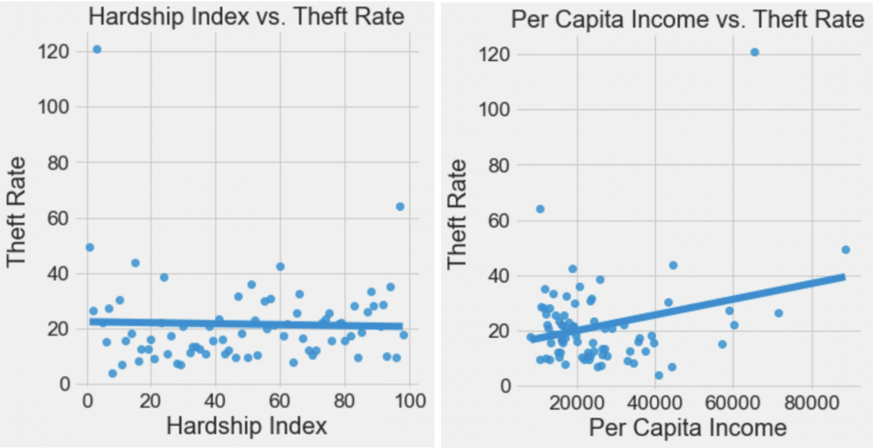
Figure 3



Demographic Variables	Standardized Coefficients
PERCENT OF HOUSING CROWDED	-5.445834
PERCENT HOUSEHOLDS BELOW POVERTY	-1.947245
PERCENT AGED 16+ UNEMPLOYED	4.444897
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	-5.533439
PERCENT AGED UNDER 18 OR OVER 64	-9.810351
PER CAPITA INCOME	10.335528
HARDSHIP INDEX	20.872034

We then graphed the two demographic variables with the greatest standardized coefficients (strongest predictors of theft) against the theft rate as seen in *Figure 4.* The per capita income demographic variable had a high linear correlation with theft rate of 0.275, however, the hardship index had a fairly low linear correlation coefficient of -0.032. The low correlation coefficient for hardship index is puzzling, but could be a result of a latent effect, in that all the other demographic factors conspired with each other to make the effect cancel overall. This effect could be made possible because the hardship index is made of a variety of economic and social factors, and consequently will have correlations with a variety of different features in the dataset. The demographic variable with the highest linear correlation coefficient was percent households below poverty, which has a correlation coefficient of 0.280.

Figure 4



Crime Type Classification with Machine Learning:

In addition to predicting theft rate, we also used various machine learning classification algorithms to predict the type of crime committed based on the features mentioned above. The categories of crimes that could be committed are Battery, Theft, Burglary, Robbery, Motor Vehicle Theft, Narcotics, Criminal Damage, Assault, and Deceptive Practice. The Models we used were Logistic models, Decision Trees, K Neighbors Classifiers, Random Forests, Stochastic Gradient Descent, and a neural network model. We used Scikit-Learn and TensorFlow to create these models. After testing these models, we determined that the Logistic, decision tree, and random forest models all performed similarly with an accuracy of ~ 40%. As a measure of effectiveness, we calculated that an accuracy of ~ 25.8% would result from classifying all crimes as the most frequent category, Theft, and an accuracy of 11% would result from randomly choosing one of the nine categories. Therefore, our achieved accuracy of ~ 40% was better than random choice, however could be improved upon.

Optimization:

Our next step to improve the model was optimization. First, we used a grid search algorithm to test multiple hyperparameters. We tested C values of .001, .009, .01, .09, 1, and 3 paired with a penalty of none or l2/ridge. After running the Grid search we determined that the best C value was .001 and the best penalty was l2. These hyperparameters improved the accuracy to ~ 43.4%. We also ran a randomized search but it performed worse.

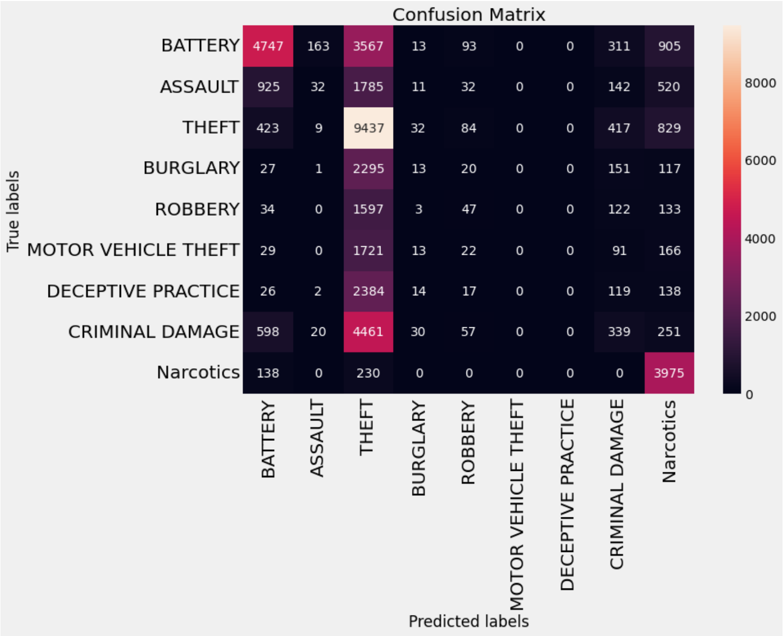
Final Evaluation:

Our Final Logistic Classification model performed with an accuracy of ~ 43.4% on the test data and ~ 43.3% on the training data. Additionally the model had a precision of 33%. The accuracy score is calculated by # of true classifications divided by # of all classifications and precision is calculated by # of true positives divided by (# of true positives + # of false positives). *Figure 5* shows a table with the different models we tested and their accuracy and precisions. Once again, ~ 25.8% would result from classifying all crimes as the most frequent occurrence. Therefore, our model performed ~17.6% better than this. *Figure 6* shows a confusion matrix with how the model predicted the crime type vs what the true crime types were. The diagonal going from top left to bottom right represents the correctly predicted labels. The brighter colors represent categories that were predicted more frequently. There are only 3 columns that were predicted significantly more correctly than the rest. These commonly predicted crimes are theft, battery, and narcotics. These crimes were also the top four most frequent crimes in the dataset. Criminal damage was the 3rd most frequent in the dataset, however, it was not predicted frequently. There are also two columns, motor vehicle theft and Deceptive practice with no predictions at all. Part of the reason why they were not predicted could be because they were less frequent in the dataset. However, this cannot be the only reason because the least frequent column, Robbery, still had predictions. Overall, even after optimization, our model was still not completely successful, however, we still were able to predict the type of crime with an accuracy ~17.6% greater than the most frequent crime. We believe our model was unsuccessful because of the data we fed it. Although our dataset had over 200k rows, there were only 56 unique values for the economic information we implemented in our model. Therefore, our model did not have enough information to predict certain types of crime.

Figure 5

Model	**Acc**	**Weighted Average Precision**
Logistic	.43	.33
Decision Tree	.43	.33
KNN	.35	.31
Random Forest	.43	.33
SGD	.42	.32

Figure 6



Interpretation/Significance of our Results: *Figure 7* shows the top eight predictor columns of crime type and the absolute value of their weights. It is important to realize that all the features depend on each other, however, we only included the top ten. By analyzing the weights from this table and generating graphs to make specific trends more apparent we will explain why certain trends might be occurring by consulting outside research.

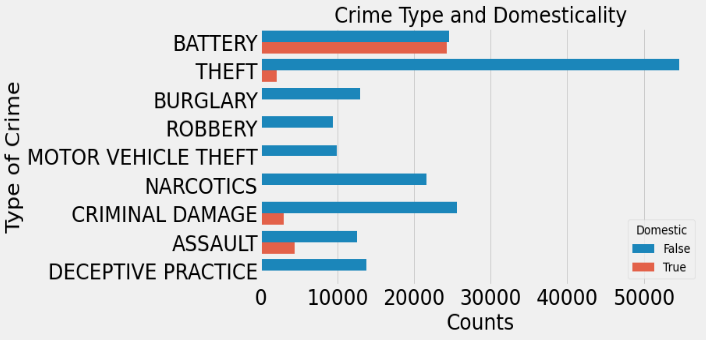
Figure 7

Feature	Weight
Domestic_False	0.456993
Domestic_True	0.456993
HARDSHIP INDEX	0.234886
PERCENT AGED UNDER 18 OR OVER 64	0.104859
PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	0.084606
PERCENT HOUSEHOLDS BELOW POVERTY	0.079443
PERCENT AGED 16+ UNEMPLOYED	0.057646
PER CAPITA INCOME	0.042998

Domesticality:

The first two rows in *Figure 7* represent if the crime was committed domestically or not. This contributed the highest weights, ~ .46, to our logistic model. It makes sense that domesticality is a high weight to our model because certain crimes should be less likely to occur inside one's own household. For example, *Figure 8* shows the types of crimes we analyzed and if they were committed domestically or not. It makes sense that burglaries, robberies, and deceptive practices do not occur domestically, because it would not make sense to rob or scam your own household. What was surprising, however, was that Battery and Assault was a common occurrence domestically. Almost half of all battery charges were domestic. These results are significant because by recognizing that battery occurs frequently in households, authorities can provide family and couple counseling services to prevent future domestic battery cases. This proved to be effective in a study conducted by Hamadan University of Medical Sciences where they concluded that “Family-based counseling intervention reduced the mean score of domestic violence in the intervention group from 68.58 ± 9.21 before the intervention to 49.56 ± 8.83 after intervention” [8].

Figure 8



Hardship Index:

The third row in *Figure 7* represents the hardship index of the community the crime was committed in. This measure was the next most meaningful contribution to our logistic model with its weight of .234. The HSI is calculated by analyzing a range of economic factors that contribute to hardship. In general, a higher hardship index score indicates a greater level of economic and social hardship. *Figure 9* shows a visualization of the HSI score vs the type of crime committed. The HSI was on average the lowest for theft and deceptive practice. Whereas, the average hardship score was the highest for narcotics as seen in *Figure 9.* These results make sense because neighborhoods with low HSI contain wealthier people, and scamming/robbing wealthier people presents a greater potential gain than scamming poor people. Therefore, the median HSI scores for theft and deceptive practices are lowest because these activities happen in wealthy areas. On the other side, crimes like narcotics might happen in areas with high HSI scores because of a higher concentration of gang related activities in poorer neighborhoods. “In poor areas in the inner cities, getting rich by selling drugs became the standard way to get ahead”[7].

Figure 9

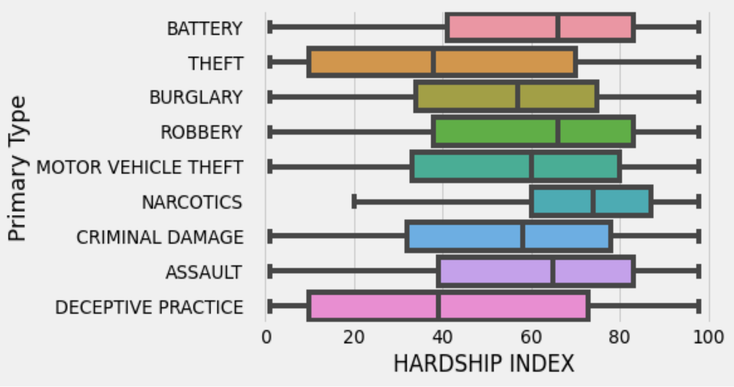
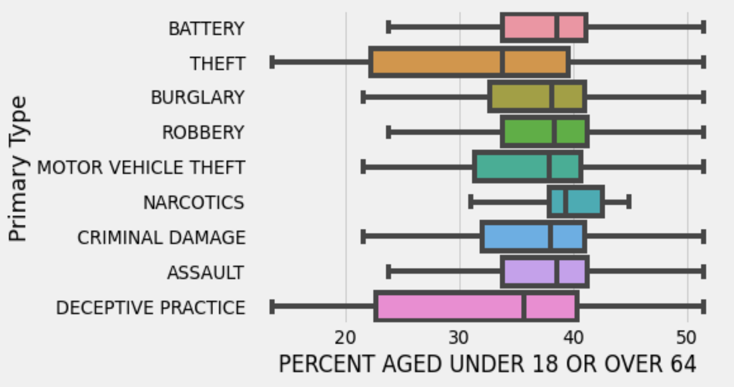


Figure 10



Percent Aged under 18 or over 64:

The last feature we will examine in depth is Percent Aged under 18 or over 64 (row 4 in *Figure 7*). This column attributed a weight of .1 to our logistic model. It is reasonable to assume that age has an effect on the type of crime committed because “Neuroscience suggests that the parts of the brain that govern risk and reward are not fully developed until age 25, after which lawbreaking drops off. Young people are more likely to be poor than older people, and poorer people are more likely to commit crimes” [9]. *Figure 10* shows that more severe crimes like battery, robbery, and criminal damage have higher percentages of people under 18 and over 64. Even though we do see a correlation between this, we realize that it would be better to separate the percentage of people under 18 and over 64 into two different features, however, the dataset we used did not make this possible.

Conclusion:

As stated earlier, the results of our logistic classification model were not successful. Out of the 9 categories of crimes we were trying to classify, only 3 of them could be classified somewhat correctly. However, from analyzing the weights of each feature in our model, we were still able to generate visualizations that demonstrated specific trends for our features vs crime type. Furthermore, by analyzing these trends, we were able to consult outside research to deduce why our features may correlate to such crimes.

In addition to the classification model, the results of our multiple linear regression for theft rate detailed that the hardship index, per capita income, percent aged under 18 and over 64, and percent aged 25+ without a high school diploma were the four strongest predictors of theft. These results could provide governments with the data to influence their policies on crime reduction in their jurisdiction: decreasing the hardship index and increasing educational level in each community area will help reduce the crime rate, and knowing that areas with higher per capita income experience greater theft can give them a better sense of what areas to monitor for long term theft prevention. This regression model could also be used to predict other types of crime in the dataset, for a better understanding of the different socioeconomic factors that influence each.

In the future, researchers interested in crime data can combine classification algorithms to classify crime type with regression algorithms to predict specific crime rates in order to predict how often and what types of crimes will occur. This information would be crucial for police departments as well as public policy and law makers.

Noah Bhuta, Stella El-Fishawy, Alena Zeng
DATA 11900

References:

[1] Chicago Police Department. “Crimes - 2015: City of Chicago: Data Portal.” *Chicago Data Portal*, 2015 (last updated February 13, 2023), https://data.cityofchicago.org/Public-Safety/Crimes-2015/vwwp-7yr9.

[2] “Below Poverty Level by Community: City of Chicago: Data Portal.” *Chicago Data Portal*, US Census Bureau, 4 June 2013, https://data.cityofchicago.org/Health-Human-Services/below-poverty-level-by-community/b7zw-zvm2.

[3] Farooqui, Suniya, “Chicago Community Area Indicators, 2015.” *Heartland Alliance*, Feb. 2017, https://www.heartlandalliance.org/povertyreport/wp-content/uploads/sites/26/2017/02/PR17_DATABOOK_CCA.pdf.

[4] Ross, Marisa. “Dr. Marisa Ross: Using Social Network Science to Prevent Gun Violence in Chicago.” *Everytown Research & Policy*, 3 Nov. 2022, https://everytownresearch.org/dr-marisa-ross-using-social-network-science-to-prevent-gun-violence-in-chicago/.

[5] Rosser, Gabriel, et al. “Predictive Crime Mapping: Arbitrary Grids or Street Networks?” *Journal of Quantitative Criminology*, vol. 33, no. 3, 2016, pp. 569–594, https://doi.org/10.1007/s10940-016-9321-x.

[6] Rotaru, Victor, et al. “Event-Level Prediction of Urban Crime Reveals a Signature of Enforcement Bias in US Cities.” *Nature Human Behaviour*, vol. 6, no. 8, 2022, pp. 1056–1068., https://doi.org/10.1038/s41562-022-01372-0.

[6] Baughman, Shima. “How Effective Are Police? The Problem of Clearance Rates and Criminal Accountability” *Social Science Research Network,* 2 Apr. 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3566383.

[7] No byline. ”Gangs and Drugs .” Drugs, Alcohol, and Tobacco: Learning About Addictive Behavior. *.Encyclopedia.com.* 14 Feb. 2023 https://www.encyclopedia.com/education/applied-and-social-sciences-magazines/gangs-and-drugs.

[8] Babaheidarian, F., Masoumi, S.Z., Sangestani, G. *et al.* “The effect of family-based counseling on domestic violence in pregnant women referring to health centers in Sahneh city, Iran, 2018.” *Ann Gen Psychiatry* 05 Jan. 2021 https://annals-general-psychiatry.biomedcentral.com/articles/10.1186/s12991-021-00332-8#citeas.

[9] Goldstein, Dana. “Too Old to Commit a Crime” *The Marshall Project,* 20 Mar. 2015, https://www.themarshallproject.org/2015/03/20/too-old-to-commit-crime.